CHAPTER 8
SCORING

The intent of this chapter is to give an overview of the procedures used to score the statewide assessment administered in Kentucky. Included are tables containing information concerning the reliability of the scoring system.

OPEN-RESPONSE QUESTIONS AND ON-DEMAND WRITING

During the four school years contained in Cycle 3 of KIRIS, open-response questions and On-Demand Writing prompts at grades 4/5, 7/8, and 11/12 required hand scoring by the contractor. Portfolios in writing and mathematics (mathematics until 1997), as well as alternate portfolios, were hand-scored by Kentucky teachers. Performance events and mathematics portfolios administered in 1994 and 1995 were not used in the Cycle 3 accountability system. The processes of selecting and training scorers, reading and scoring papers, and monitoring scoring results remained similar to those carried out in previous years. The 1995 and 1996 years testing constituted the baseline for Cycle 3, and are described in detail in the *KIRIS Accountability Cycle 2 Technical Manual*. Modifications in training materials for scorers and teachers are described as well.

The primary contractor changed following the discovery of errors in the reporting of 1996 scores. Data Recognition Corporation (DRC) of Minnetonka, Minnesota, was contracted for the hand scoring of student open-responses in 1997 and 1998. Scoring Guide development was a cooperative effort involving KDE, WestEd, and DRC. Student response booklets were scored immediately following login. The following material describes the scoring procedures implemented during the final two years of Cycle 3.

SELECTION OF SCORER EQUATING PAPERS. KIRIS used matrix sampling within the content areas of the test. This allowed for a larger number of questions that evaluated a larger number of specific content statements from the *Core Content for Assessment*. By means of matrix sampling questions selected by KDE remained in the test for two to four years. This allowed different students in different years to answer the same questions. This allowed KDE to determine whether students (and hence, the schools) were improving their performance over the years. Since the same people were not scoring the same questions from year to year, procedures had to be in place to ensure the scorers maintained consistency in evaluating responses.

One method of assessing consistency across scoring groups is by selecting and integrating into the scoring process responses scored the prior year with results of prior years unknown to current scorers. Whether the scorer knows that the item is from a prior year is irrelevant since the expectation is that they score as they have been trained to score the current year's items.

To accomplish this consistency check, a four-step process is used.

Step 1 - During test development, 10 to 15 questions are selected as scorer equating questions.
Step 2 - A random sample of 50 responses at each score point for each equating question is generated and pulled from the previous year's files.
Step 3 - These responses are distributed and scored according to the current year's guidelines.
Step 4 - The score sheets are scanned and a separate file maintained for use in determining scorer accuracy.

If analysis reveals a difference in scores between years, a mathematical adjustment to scores can be applied to maintain consistent scores over the years. (See Tables 9-3a, 9-3b, and 9-3c for intercept adjustments for scorer differences)

SELECTION OF RANGEFINDING PAPERS. Prior to scorer training, student responses are scored during a process called range finding. This scoring is a joint effort of DRC scoring staff, KDE Curriculum Development staff, and WestEd test development personnel. Rangefinding is used to judge the validity of the scoring guides for each item, that is, to ascertain if students responded as expected by the guide writers. The second use of rangefinding is for selecting responses at each score point for training scorers.

The student responses used in rangefinding come from the first papers received from local districts. Following the scoring of these initial items, the scoring guides are finalized. With finalized guides and a set of scored responses, the materials needed for the training of scorers are compiled. Responses that are particularly important are annotated for use as anchor papers that illustrate the score points in the scoring guide for each item. Additional responses are selected and scored by DRC, KDE and WestEd personnel for use in packets. Qualification packets are used for establishing scorer readiness to begin scoring.

STAFFING. Levels of staffing for Cycle 3 are listed in Table 8–1. The table also shows the percentages of scorers at each grade level who participated in a previous year's scoring (repeat scorers), as well as the number of training leaders. The comparatively low percentage of repeat scorers in 1996, particularly at grade 8, is due in part to opening a second scoring center to handle several forms of the test for grades 4 and 8. The low percentage of repeat scorers in 1997 is the result of the transition from the prior contractor to DRC.

QUALIFICATIONS. Table 8–2 shows education level and demographic information for scorers for Cycle 3. Qualifications for grade 11 in 1997 are not included because, by agreement, the prior contractor still scored those responses, and that information is not available.

DRC TEAM LEADER TRAINING. Comprehensive team leader training lasted three days. The scoring directors for each content area managed the training. Team leader training followed the procedures used in scorer training but was more comprehensive to accommodate the responsibilities required of team leaders. During their training, team leaders were required to annotate all their training responses with official KDE/DRC annotations. It was important that each team leader impart the same rationale for each response to promote room wide scoring consistency.

SCORER TRAINING. During the baseline years of 1995 and 1996, the prior contractor completed the training of scoring staff. A full explanation of those procedures is available in the *KIRIS Accountability Cycle 2 Technical Manual,* Chapter 8

TABLE 8–1
NUMBER OF SCORERS AND TRAINING LEADERS AT EACH GRADE

Grade	1995			1996[1]			1997			1998		
	% Repeat Scorers	Scorers	Training Leaders	% Repeat Scorers	Scorers	Training Leaders	% Repeat Scorers[2]	Scorers	Training Leaders	% Repeat Scorers	Scorers	Training Leaders
4/5	87	203	14	63	267	21	0	176	19	69	160	17
7/8	83	263	17	9	221	20	0	192	21	75	176	19
11	67	172	14	73	167	25	0	N/A	N/A	N/A	159	17

[1] The comparatively low percentage of repeat scorers in 1996, particularly at grade 8, is due in part to opening a second scoring center to handle several forms of the test for grades 4 and 8.

[2] The low percentage of repeat scorers in 1997 is the result of the transition from the prior contractor to DRC.

TABLE 8–2
PROFILE OF SCORER QUALIFICATIONS AND DEMOGRAPHICS

Background		Number of Scorers											
		1995			1996			1997			1998		
		Grade 4	Grade 8	Grade 11	Grade 4	Grade 8	Grade 11	Grade 4/5	Grade 7/8	Grade 11[1]	Grade 4/5	Grade 7/8	Grade 11
Education	Degrees beyond Bachelor's Degree	35	52	25	52	18	42	29	36	-	28	31	27
	Bachelor's Degree	110	156	119	159	78	104	134	141	-	115	120	111
	Associate's Degree	4	5	4	9	19	8	5	3	-	6	9	6
	Two-year college study or equivalent	54	50	24	47	106	13	8	12	-	11	16	15
Demo-graphics	Male	76	97	70	104	91	71	82	90	-	75	83	74
	Female	127	166	102	163	130	96	94	102	-	85	93	85
	Black	2	4	2	3	6	3	5	5	-	3	6	5
	White	201[1]	259[1]	170[1]	261	215[1]	161	165	180	-	155	165	150
	Other[2]	0	0	0	3	0	3	6	7	-	2	5	4

1. Data are not available. DRC did not score Grade 11 in 1997.
2. Data collection did not include the other category entitled “other.”

SCORER TRAINING FOR 1997 AND 1998

All DRC scorers had a minimum of two years of college study or the equivalent and scored in content areas where they had expertise, training, and/or experience. The two stages of training were organized and monitored by the DRC Hand-Scoring Project Director. The scoring directors were trained first, focusing upon Kentucky objectives, content guidelines, and the standards. The second phase was the training of the scorers. KDE and WestEd Development personnel were allowed to participate and monitor the training as agreed per contract.

Training for scorers for each content area began with a presentation of the standards and discussion of the scoring guide and anchor papers by the scoring director. Practice scoring and thorough discussions of each of the training sets in each of teams followed this presentation. The small group discussions were conducive to understanding the score point explanations.

After scorer training, scorers demonstrated their ability to apply the scoring criteria at an 80% level of agreement with true scores on two qualifying sets. Any scorer who did not qualify was retrained until able to qualify. All scorers hired for this project were able to qualify. Daily procedures maintained the level of qualification of the scorers.

PROCEDURES AND QUALITY CONTROL

Training of scorers was completed before the KIRIS responses were received in April of each year. Scoring began immediately.

Packets were distributed to each table of scorers. Packets contained fifteen Student Response Booklets (for grades 4, 5, 7, and 8) along with score sheets for two scorers per content area. The scorer pulled out the score sheet for the appropriate content area and checked the packet number and student lithocode of the score sheet against the packet header. When these were confirmed, the scorer checked the score codes in his/her scorer ID number and scored the responses. When scorer 1 was finished, he/she placed the packet in the bin on the scoring table for pick-up by a clerk. Packets that required a second reading (part of the 2% read behind procedure), were distributed by clerks. When a packet was complete, a clerk filed it and took the score sheets to be scanned.

The content areas of Reading, Mathematics, Science, and Social Studies each had seven open-response items per grade tested (4 common, 3 matrix which included 1 pre-test); Arts & Humanities and Practical Living/Vocational Studies had three matrix items (2 operational matrix and 1 pre-test) per student. To ensure that no school or student had all responses scored by the same person, each student's responses were scored by at least two scorers. In Reading, Mathematics, Science and Social Studies, one scorer scored responses A-D (all common questions) and the second scorer scored responses E and F (two matrix). The second scorer also scored the pre-test item if the student's response was included in the sample of approximately 500 pre-test responses scored. For Arts & Humanities and Practical Living/Vocational Studies only one scorer was necessary since there were only three questions per student. For 2 percent of the student papers a quality control scoring was accomplished. Two persons scored the common and matrix responses. These scorers placed the packet and score sheet in their team leaders basket for review. If a score between the original scorer and the quality control scorer differed, but were adjacent scores, the original score stood as the score of record. If the scores differed by more than one score point, the difference was resolved by the content area Scoring Director or team leader, with their score becoming the score of record.

As an additional measure in scorer quality control, scoring directors and team leaders used a read behind log to track an individual's scoring consistency. Team leaders randomly read and scored eight to ten student responses in a packet and then compared their scores with the scorer of record. If the team leader saw a pattern of errors or consistent errors, he/she retrained the scorer. The frequency of targeted read behinds decreased as the scorer demonstrated aptitude, but read behinds were always done at least once a day for each scorer.

In order to monitor scorer reliability and to ensure that an acceptable agreement rate was maintained, DRC monitored the daily statistics provided by the reliability report. The reliability report documented individual scorer data including scorer number, number of responses scored, individual score point distributions, and agreement rates for the 2 percent of the responses which received a second reading. In addition to this information, DRC used scorer statistics on individual performances on the recalibration sets to monitor scorer accuracy.

CONSISTENCY OF SCORING. Scoring of open-response items was monitored as scoring directors and team leaders constantly moved from scorer to scorer, reading examples of each scorer's work. Each team leader read approximately one packet per scorer per day.

Tables 8-3 and 8-4 contain data concerning the agreement between scorers. In each table, the first column is the percentage of student responses in the 2 percent read behind sample, where both scorers agreed exactly. The second column reflects the percentage of those responses where the scorers differed by one point. The third column indicates those where the scorers disagreed by more than one point. The tables indicate that the agreement of scorers significantly exceeded the 80 percent standard except in science in grade 4 in 1998.

TABLE 8-3
READER MONITOR REPORT (PERCENTAGES) 1997

:	Grades 4/5			Grades 7/8			Grade 11[1]		
Content	Exact	Adjacent	Non-Adjacent	Exact	Adjacent	Non-Adjacent	Exact	Adjacent	Non-Adjacent
Reading	84	16	0	86	14	0			
Math	91	9	0	89	11	0			
Science	80	19	1	83	16	1			
Social Studies	91	10	0	82	17	1			
Arts & Humanities	88	12	0	91	10	0			
PL/VS	89	11	0	90	10	0			
Writing	90	10	0	90	10	0			
Total	87.4	12.4	0.1	87.1	12.5	0.3			

[1] Data are not available. DRC did not score the 1997 open-response questions at Grade 11.

NOTE: Percentages may not total 100% due to rounding.

TABLE 8-4
READER MONITOR REPORT (PERCENTAGES) 1998

	Grades 4/5			Grades 7/8			Grade11		
Content	Exact	Adjacent	Non-Adjacent	Exact	Adjacent	Non-Adjacent	Exact	Adjacent	Non-Adjacent
Reading	84	16	0	88	12	0	86	13	1
Math	88	12	0	87	13	0	89	10	1
Science	77	21	2	82	17	1	87	13	0
Social Studies	85	15	0	88	12	0	91	9	0
Arts & Humanities	91	9	0	90	10	0	90	10	0
PL/VS	90	10	0	89	11	0	89	11	0
Writing	88	12	0	84	16	0	83	17	0
Total	86.1	13.5	0.3	86.8	13.0	0.1	87.8	11.8	0.3

NOTE: Percentages may not total 100% due to rounding.

SCORING SHEETS

The scoring sheets used for the open-response questions were specially designed to meet the requirements of the scoring process. Scoring sheets for each grade level were color-coded and the content area, question/page number, and student lithocode were preprinted on the form. There was one scoring sheet to capture scores for the common items and one scoring sheet for the matrix and pretest items. The scoring sheets contained places for the scorer to code his/her ID number and scoring grids (0, 1, 2, 3, 4, and B for blank) for each student's responses. Duplicate (Scorer 2) scoring sheets were generated for packets that were part of the 2 percent read behind.

DRC was required to score at least 500 responses for each of the twelve pre-test questions per content area. The scoring sheet was marked to identify the first and seventh student as needing the pretest question scored in 4, 5, 7, and 8. The first student in every packet in grade 11 had his/her response to the pre-test questions scored.

PERFORMANCE EVENTS (1995 only)

Scoring methods for the Performance Events were described in the *KIRIS Accountability Cycle 2 Technical Manual.* Performance Event scores were not included in the Accountability Cycle 3 system and are not considered further here.

PORTFOLIO SCORING

The scoring of Writing and Alternative Portfolios by Kentucky teachers and the audit process that assures quality in the scoring processes are described in Chapter 12. The Mathematics Portfolio, which was administered in 1995 and 1996, but not in 1997 or in 1998, was not part of the accountability system in Cycle 3. See Chapter 12 for a complete discussion of the Mathematics Portfolio.

TABLE 8-5
INTER-RATER RELIABILITY IN SCORING OF OPEN-RESPONSE QUESTIONS
Percentage of Exact Agreement

Grade	1995				1996			
	Grade 4	Grade 8	Grade 11	Total	Grade 4	Grade 8	Grade 11	Total
Reading	85.7	86.2	84.8	85.6	89.5	90.3	88.6	89.5
Math	95.3	97.1	93.6	95.5	96.4	98.5	97.9	97.6
Science	87.3	85.9	86.2	86.4	86.3	82.5	80.9	82.9
Social Studies	85.0	85.2	84.0	84.8	83.3	80.0	79.3	80.8
Arts & Humanities	87.7	87.6	87.1	87.5	90.1	87.3	89.8	88.8
PL/VS	89.3	87.3	88.7	88.3	90.2	88.7	88.5	89.5
Total	87.8	87.9	86.8	87.5	88.7	86.6	85.5	86.9

TABLE 8-6
INTER-RATER RELIABILITY IN SCORING OF OPEN-RESPONSE QUESTIONS
Percentage of Within 1 Score Point

Grade	1995				1996			
	Grade 4	Grade 8	Grade 11	Total	Grade 4	Grade 8	Grade 11	Total
Reading	99.8	99.9	99.8	99.8	99.8	99.9	99.8	99.8
Math	99.9	99.9	99.8	99.9	99.4	99.7	99.7	99.6
Science	99.8	99.8	99.7	99.8	99.4	99.0	99.1	99.2
Social Studies	99.9	99.9	99.9	99.9	99.8	99.5	99.6	99.6
Arts & Humanities	99.8	99.9	99.7	99.8	99.9	99.8	99.9	99.8
PL/VS	100.0	100.0	99.9	100.0	99.9	99.9	99.9	99.9
Total	99.9	99.9	99.8	99.9	99.7	99.5	99.6	99.6

TABLE 8-7
STUDENT-LEVEL INTER-RATER CORRELATIONS

Subject	1995			1996		
	Grade 4	Grade 8	Grade 11	Grade 4	Grade 8	Grade 11
Reading	.92	.88	.90	.93	.93	.92
Math	.98	.99	.98	.97	.99	.98
Science	.92	.91	.92	.91	.91	.90
Social Studies	.91	.91	.92	.90	.87	.88
Arts & Humanities	.91	.91	.90	.94	.88	.91
PL/VS	.92	.91	.91	.92	.91	.90

CONCLUSION

The descriptions and statistics contained in this chapter indicate the intensity of the effort to assure that open-response items were scored as accurately as possible. Every possible means was used to prevent scorer drift toward easier or harder scoring.

This page was intentionally left blank.